低ランク近似を介した選択的パラメータ更新による差分プライベート学習

伊藤　竜一† リュウセンペイ†† 高橋　翼†† 佐々木勇和† 鬼塚　真†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1-5
†† LINE 株式会社　〒 160–0004 東京都新宿区四谷 1-6-1 四谷タワー 23 階
E-mail: †{ito.ryuichi,sasaki,onizuka}@ist.osaka-u.ac.jp, ††{sengpei.liew,tsubasa.takahashi}@linecorp.com

あらまし　パーソナルデータから機械学習モデルを学習する際に，差分プライバシにそった厳密なプライバシを保証する訓練手法が提唱されている．広く知られている DP-SGD という手法では，勾配クリッピングやノイズ付与によってプライバシ保護を行う一方，結果として得られるモデルの有用性は低下してしまう．本稿では，モデルの重み行列を低ランク近似した上でスパース性を導入することで，DP-SGD による差分プライバシと高い有用性を両立した訓練手法を提案する．自然言語処理のベンチマークを利用し，提案手法がプライバシ強度を落とすことなく有用性を改善することを確認した．

キーワード　差分プライバシ，深層学習，低ランク性，スパース性

1　はじめに

様々なデータの蓄積が容易になり，個人に関するデータ，つまりパーソナルデータの利活用が進んでいる．パーソナルデータの利用はサービスに有用であることが多いが，一方で個々人の特定やプライバシの開示を防ぐためプライバシの保護が必要となる．よく知られたプライバシ保護の方法としてデータの k-匿名化が挙げられる．しかしながら，リンケージ攻撃以外の攻撃や複数の問合せによるモザイク効果などによってプライバシが開示されてしまうこと知られており [1]，プライバシ保護として厳密とは言えない．このような背景から，攻撃手法や背景知識に依存しないプライバシ指標として差分プライバシ [2] が提案されている．差分プライバシは一定の識別困難性に基づき任意の攻撃に対してプライバシ保護を行う．例えばある計算機構が差分プライバシを満たしている場合，いかなる攻撃手段や背景知識を利用してもプライバシ強度パラメータで指定される以上に特定レコードを識別することが不可能と保証される．差分プライバシの有用性は既に認知され始めており，例えば Apple 社はユーザデータを差分プライバシで保護して扱っていることを公表している [3].

データを元に予測や推定を行う機械学習でもパーソナルデータが含まれる場合はプライバシ保護が必要となる．差分プライバシも選択肢の 1 つであり，学習時に差分プライバシを満たす，差分プライベート学習手法が提案されている．ニューラルネットワークでは，一般的に，確率的勾配降下法（SGD）によってパラメータの更新を行うため，パラメータにパーソナルデータの情報が埋め込まれることとなる．Abadi ら [4] はこの点に注目し，SGD によるパラメータ更新の際に勾配クリッピングとノイズ付与を行うことで差分プライバシを満たす Differentially Private Stochastic Gradient Descent (DP-SGD) を提案している．DP-SGD は安全性を向上させる一方で，SGD と比較して，勾配クリッピングとノイズ付与の影響によりモデルの有用性が低下するというトレードオフがある．そのため，タスクや安全性の度合いによっては有用なモデルが得られないことがある．差分プライバシを満たすニューラルネットワークモデルの実用上の意義は大きく，DP-SGD の有用性が低下するという問題を解決する手法も提案されている [5, 6, 7]．いずれの手法も更新するパラメータを減らすことでクリッピングやノイズの影響を抑えるというアプローチをとる．特に重み行列や勾配行列を低ランク近似して扱うことでモデルの有用性を向上できると報告されている [5, 6]．しかしながら依然として，プライバシを考慮しないモデルと比較すると有用性は低くなっている．

本稿では，ニューラルネットワークパラメータの低ランク性とスパース性に基づき更新対象パラメータを適切に選択することで，差分プライバシを満たしながら有用性の高いモデルとして学習する手法を提案する．学習への寄与が少ないパラメータを更新から除外することで，差分プライバシのための勾配クリッピングとノイズ付与の影響を減らし，最終的に得られるモデルの有用性の向上を図る．まず，提案手法では既存手法を用いて低ランク性に基づく更新対象パラメータ削減を行う．低ランク性による更新対象パラメータ削減手法はいくつか提案されており [5, 6]，提案手法ではそれらからタスクに応じて最も適切なものを選択できる．これに加えてスパース性による更新対象パラメータ削減を導入する．各ニューロンの重要度を集約して入力ユニットと出力ユニットに対応付けることで，低ランク化したパラメータに対するスパース化を行う．これにより低ランク性とスパース性を両立した更新対象パラメータ削減となり，有用性の高いモデルが得られる差分プライベート学習を実現する．

自然言語処理のベンチマークを用いた実験により，提案手法の有効性を評価する．ベースとなる低ランク性を利用した差分プライベート学習手法として RGP [5] と LoRA+DP-SGD [6] を利用する．言語モデル RoBERTa [8] の差分プライベートなファインチューニングを GLUE ベンチマーク [9] で評価したところ，安全性を変えることなくモデルの有用性を表す正答率を

最大で 3%向上させることに成功した．

2 章で差分プライバシとニューラルネットワークに関する事前知識を導入し，3 章で提案手法を詳説する．4 章で提案手法の評価を行い，5 章で本稿をまとめる．

2 事前準備

この章では提案手法の基礎となる概念や既存技術について説明する．

2.1 (ϵ, δ)-差分プライバシ

差分プライバシ [2] とは，データベースに含まれるパーソナルデータの保護を目的とした指標である．差分プライバシを満たした計算機構の場合，いかなるレコード集合が対象となる問合せを組み合わせても特定のレコードが含まれているかどうかの識別が困難となる．これを以て特定のレコード，つまりパーソナルデータのプライバシが開示されないことを保証している．識別困難性に基づいているため，攻撃手法や背景知識に依らない指標となっている点が特徴的である．

定義 1 ((ϵ, δ)-差分プライバシ [2])．レコード x_i の集合をデータベース $D = \{x_i\}_{i=1}^n$ とし，取り得るデータベースの集合を $\mathcal{D}$ とする．D から情報を取り出す処理を Q とし，その出力を $R = Q(D)$ ，取り得る出力の集合を $\mathcal{R}$ とする（$R \in \mathcal{R}$）．任意の隣接したデータベースの組 $(D_1, D_2) \in \mathcal{D}$ [1]に対して以下が成立するとき， Q が (ϵ, δ)-差分プライバシを満たすと言う．

$$\mathbf{Pr}[Q(D_1) \in R] \leqq e^{\epsilon} \cdot \mathbf{Pr}[Q(D_2) \in R] + \delta \tag{1}$$

任意の隣接したデータベース D_1, D_2 に対する問い合わせ結果 $Q(D_1)$ と $Q(D_2)$ を観測してもパラメータ ϵ， δ で指定された程度に識別が困難であり，特定のレコードがデータベースに含まれているかどうかの推定が困難であることを表している．

2.2 Differentially Private Stochastic Gradient Descent (DP-SGD)

DP-SGD [4] とは，差分プライバシを満たす確率的勾配降下法である．モデルの構造に依らず，パラメータ更新に利用することで差分プライバシを満たすモデルとなる．具体的には，勾配クリッピングとノイズ付与を追加で行う．サンプルごとに計算された勾配の L2 ノルムを閾値 C でクリッピングすることで，サンプルごとの影響度合いに上界を定める．また，ガウシアンノイズを勾配に加算することで，(ϵ, δ)-差分プライバシの満足を保証する．

パラメータ更新に利用する勾配に変更を加えるため，差分プライバシを考慮しない場合と比較して得られるモデルの有用性は低下する．この度合いはパラメータ ϵ， δ に依存しており，安全性を向上させると有用性が低下するトレードオフとなっている．

1： ここでの隣接とは，1 レコードのみが異なることを表す

2.3 DP-SGD を拡張した既存手法

DP-SGD では差分プライバシによる安全性が担保された一方で，有用性の低下が大きく，実用に問題が出てしまうことが実験的に示されている [5]．そこで，DP-SGD と同等の安全性を担保したまま，有用性を向上させる提案が行われている [5,6,7]．いずれの手法も DP-SGD によるクリッピングとノイズの影響を緩和することでモデルの有用性向上を図っている．主に不必要なパラメータの更新を削減することで，元の勾配が持つ情報量以上にノイズの影響を受けてしまうことを避けるアプローチが取られている．このアプローチの実現方法は，ニューラルネットワークに表れる重み行列と勾配行列の低ランク性を利用したものとスパース性を利用したものの 2 つに大別される．

2.3.1 低ランク性に基づく手法

低ランク性とその利用について紹介する．行列のランクが低いということは，ランク分解を行い要素数の少ない行列として扱っても元の行列の情報を保てることを意味する．ニューラルネットワークの重み行列と勾配行列の低ランク性は経験的に知られており [5,6,10,11,12,13]，DP-SGD の拡張以外に圧縮や高速化のためにも利用され，その有効性が示されている．

Reparametrized Gradient Perturbation (RGP)： RGP [5] とは，低ランク近似した重み行列を利用する差分プライバシを満たすニューラルネットワークのパラメータ更新手法である．重み行列 $W \in \mathbb{R}^{m \times n}$ を，低ランク近似した行列 $L \in \mathbb{R}^{m \times k}, R \in \mathbb{R}^{k \times n}$ と順伝播のみに利用する残差行列 $\tilde{W}$ を用いて以下のように再定義する．

$$W \to LR + \tilde{W}.\text{stop\_gradients}() \tag{2}$$

stop_gradients() はパラメータ更新を行わないことを示す．入力を x，出力 y としたとき，順伝播は以下のように表される．

$$y = LRx + \tilde{W}x \tag{3}$$

また，L, R の勾配はそれぞれ $\partial L = (\partial W)R^T, \partial R = L^T(\partial W)$ であるため，L の列と R の行が正規直交基底を成している場合，逆伝播として W を $\partial W = (\partial L)R + L(\partial R) - LL^T(\partial L)R$ で更新できる．このとき，DP-SGD と同様に勾配クリッピングとノイズ付与を L, R に行うことで差分プライバシを満たす．DP-SGD を直接利用した場合と比較してモデルの有用性が高くなることが報告されている．

LoRA + DP-SGD： LoRA [14] とは，低ランク性に基づくニューラルネットワークモデルの効率的なファインチューニング手法である．学習済みパラメータ $W_0 \in \mathbb{R}^{m \times n}$ に追加の更新を行う一般的なファインチューニングとは異なり，W_0 には更新を行わず追加学習用パラメータ $L \in \mathbb{R}^{m \times k}, R \in \mathbb{R}^{k \times n}$ に更新を行う．L, R は W_0 をランク k で分解したものと対応があると見なせる．入力を x，出力 y としたとき，順伝播は以下のように表される．

$$y = LRx + W_0x \tag{4}$$

LoRA 自体は差分プライバシと無関係の手法だが，DP-SGD と組み合わせることで，DP-SGD を単独で利用した場合と比較してモデルの有用性が高くなることが報告されている [6]．

2.3.2 スパース性に基づく手法

スパース性とその利用について紹介する．行列がスパースであるということは，0（ごく小さい値が無視できる文脈ではそれも含む [2]）である要素が多く，効率的な格納方式や計算方式を採用できることに繋がる．多くのニューラルネットワークの重み行列や勾配行列にスパース性は表れており，圧縮や高速化を目的とした枝刈り技術を中心に広く活用されている [15,16,17]．近年では，大規模モデルに表れる重み行列のスパース性を利用した宝くじ仮説 [18] が注目されている．宝くじ仮説の詳細は 2.4 節で述べる．

Sparse Network Finetuning with DP-SGD (SNF-DP-SGD)： SNF-DP-SGD [7] とは，スパース性に基づくニューラルネットワークモデルの効率的な差分プライベートファインチューニング手法である．スパース性を仮定した枝刈り技術 [17,18] に倣い，重みの絶対値が小さいパラメータを重要でないパラメータとして扱い，そのパラメータを DP-SGD による更新対象から除外する．このときパブリックデータで事前学習した重みを参照することで，追加のプライバシコストは必要としない．一方で事前学習した重みは不変であるため，更新されるパラメータはファインチューニングの中で常に一定である．また，ドメイン適応の先行研究 [19] に基づき，スパース性の仮定は畳み込み層のパラメータのみを対象としている．DP-SGD でファインチューニングした場合と比較してモデルの有用性が高くなることが報告されている．

2.4 宝くじ仮説

宝くじ仮説 [18] とは，ニューラルネットワークモデルには有用性が元のネットワークに匹敵するスパースな部分ネットワークが存在するという仮説である．パラメータ数が多いニューラルネットワークモデルの性能が高くなりやすいのは，内包するパラメータの組み合わせから成る部分ネットワーク（くじ引き券）の数が多く，ここに当たりくじとなる有用性の高い部分ネットワークが含まれやすいためとされている．なお一般的な枝刈り技術とは異なり，部分ネットワークはその構造だけでなく初期値も当たりくじかどうかに影響していると分析されている．この仮説をもとに，反復枝刈りによる当たりくじ探索手法が提案されており，大規模ニューラルネットワークモデルの圧縮や有用性の向上への貢献が実験的に示されている．

また，後続の研究 [20,21] により，モデルが学習済みでファインチューニングの段階であってもこの仮説は成り立ち，当たりくじが存在していることが報告されている．

3 提案手法

得られるモデルの有用性が高い差分プライベート学習手法を提案する．既存の DP-SGD を拡張した手法では，ニューラルネットワークパラメータの低ランク性とスパース性のいずれかを利用して更新対象パラメータを削減することで，DP-SGD 処理の影響の軽減から有用性の改善を図っている．低ランク性とスパース性は独立したものであるが，ランク分解された行列の成分は元の行列の複数の成分に影響を持つため，その組み合わせを単純に扱うことは出来ない．そのため，低ランク性とスパース性を両立した更新対象パラメータ数削減手法は存在していない．

2：$W_{i'j'} < \theta, \theta \approx 0$ はアクティベーションへの貢献がごく小さい（$\sum^{i \neq i' \wedge j \neq j'} W_{ij}x_{ij} \approx \sum W_{ij}x_{ij}$）ため $W_{i'j'} = 0$ として扱うことができる

```
// 低ランク性のみ
0.9727, 0.9805, 1.0000, 1.0000, 0.9678, 1.0000, 1.0000,
0.9023, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000,
0.6670, 1.0000, 0.9531, 1.0000, 1.0000, 1.0000, 0.9473

// 低ランク性 + スパース性（p = 0.3）
1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000,
1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000,
0.7300, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000
```

Listing 1: スパース化によるクリッピングへの影響例．
クリッピング前後のノルム比であり，値が小さい要素ほど勾配が大きくクリップされる．

提案手法では，重み行列から各ニューロンの重要度を集約して入力ユニットと出力ユニットの重要度を定義する．低ランク分解した行列にも入力ユニットや出力ユニットとの対応関係があるため，定義した重要度を利用することで低ランク分解された勾配行列のスパース化が可能となる．これにより，低ランク性とスパース性のいずれかだけを利用した場合と比較して，両方の性質から捉えられる多くの更新対象パラメータを削減し，結果として差分プライバシのための勾配クリッピングやノイズの影響が軽減され，得られるモデルの有用性が向上する．実際に，低ランク性のみの場合と提案手法による低ランク性とスパース性を併用した場合をクリッピングの観点で比較すると，Listing 1 のようになる．各要素が学習データに対応しており，数値が 1 であることは勾配行列のノルムがクリッピング閾値を下回りクリップされないことを，数値が 1 未満である場合は値が小さいほど勾配の値が大きくクリップされることを意味する．スパース化すると勾配行列のノルムは単調に小さくなり，多くの要素でクリッピングの影響がなくなるか軽減されていることがわかる．提案手法の主な流れは以下の通りである．

（1） 重み行列から入力側ユニットと出力側ユニットの重要度を算出

（2） 低ランク分解された勾配行列の取得

（3） 勾配行列に DP-SGD による差分プライバシ処理を適用

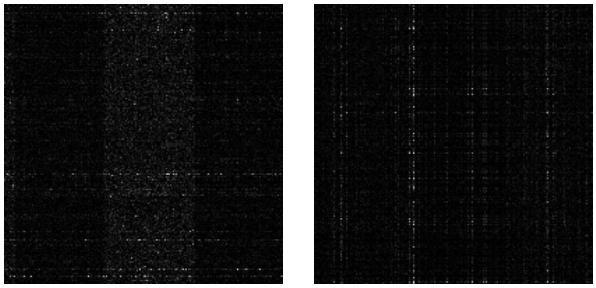
（4） 入力側ユニットと出力側ユニットの重要度に基づき勾配行列をスパース化

（5） 勾配行列を利用して重みを更新

3.1 節で更新対象パラメータ数の削減に利用する性質を個別に導入した後，3.2 節でそれらの性質の併用と具体的な手順について詳説する．

3.1 更新対象パラメータ数の削減

ここでは提案手法で利用する更新対象パラメータ数削減のアプローチについて述べる．提案手法では重み行列と勾配行列の低ランク性とスパース性の両方に注目する．例えば学習済み言語モデルである RoBERTa [8] の重み行列は図 1(a) のようになっており，多くのパラメータが ≈ 0 でスパース [2] となっていることがわかる．各軸が入力側ユニットと出力側ユニットに対応していることを踏まえると，ユニット単位に重要度の偏りがあることも確認できる．また，各軸方向に似たような傾向を持つベクトルが多数見られることから，低ランク性も示唆されている．学習フェーズに見られる勾配行列も同様の傾向であることが図 1(b) からわかる．ここからは提案手法で利用する低ランク分解とスパース化それぞれ個別に導入を行う．



(a) 重み行列　　(b) 勾配行列

図 1　学習済み RoBERTa モデルのパラメータ例．
明るい要素ほど値が大きいことを示す．

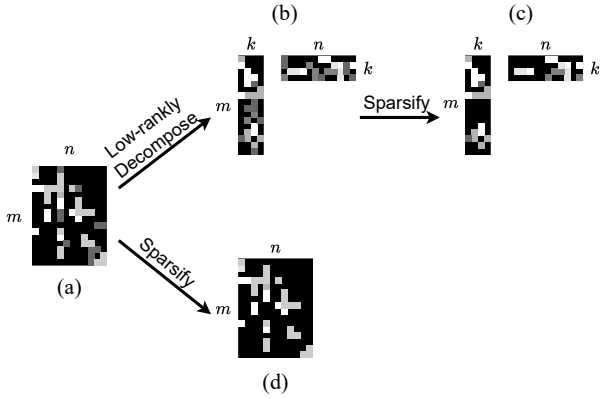


図 2　低ランク化とスパース化

まずニューラルネットワークパラメータの低ランク性の利用について述べる．この低ランク性は経験的に知られているものであり [10,11,12,13]，例えばニューラルネットワークの圧縮手法として利用されている [11,12,13]．行列 $W \in \mathbb{R}^{m \times n}$ に対してランク k での近似を考えるとパラメータ数は $k(m+n)$ となる．$k \ll \min(m,n)$ である場合 $k(m+n) \ll mn$ となり，大幅にパラメータ数を削減できることがわかる（図 2(a)/図 2(b)）．ニューラルネットワークのパラメータ更新のための低ランク化された勾配行列を得る手法には様々なものが考えられるが，提案手法は汎用的なインターフェースを備えているため具体的な手法の種類は問わない．4 章の評価実験では RGP [5] と LoRA [14] を用いているが，その他の手法でも容易に適用可能である．低ランク化した行列を扱う際の一般的な課題として，ランク k を小さくするとパラメータ数をより削減できるが，一方で復元した際の精度が低下するというトレードオフがある．提案手法では後述するスパース性と組み合わせることで削減するパラメータ数と精度の両立を行う．詳しくは 3.2 節で述べる．

次にニューラルネットワークパラメータのスパース性の利用について述べる．このスパース性も低ランク性と同じく経験的に知られており，様々な手法の前提となっている．代表的な応用としては圧縮や高速化を目的とした枝刈り手法 [15,16,17] が挙げられる．例えば $p\%$ のスパース性を仮定した枝刈りを行うと単純に $p\%$ のパラメータ数削減となる（図 2(a)/図 2(d)）．また，特に大規模モデルにおけるスパース性は宝くじ仮説 [18] を通して再確認されている．これは，ニューラルネットワークモデルで性能への寄与が大きいのは構造と初期値の組み合わせとして確率的に存在するスパースな部分ネットワーク（当たりくじ）であり，大規模モデルであることはこの当たりくじが含まれる可能性が高くなるからである，という仮説である．つまりパラメータ数が過剰な場合，ニューラルネットワークモデルの本質はスパースな部分モデルに集中しているとも言える．このことはモデルをスクラッチから学習した場合だけでなく，学習済みモデルに対するファインチューニングでも同様に成り立つと報告されている [20,21]．提案手法では重みの絶対値を重要度としたスパース性を利用する．絶対値を利用するというアプローチは，重みは正負に依らず貢献があり，絶対値が小さい重みはアクティベーションに対する貢献が少ないということに基づいている．単純ながら多数の枝刈り手法で有効性が示されている [15,16,17]．加えて，一部の枝刈り手法でも利用されているユニット単位の重要度 [16] を導入する．入力側ユニットと出力側ユニットのそれぞれで，接続されているシナプスの重要度の総和をユニットごとに計算し，それを各ユニットの重要度とする．i 番目の入力側ユニットの重要度 I_i と j 番目の出力側ユニットの重要度 O_j は以下のように算出される．

$$I_i = \sum_{j=1}^{n} |W_{ij}| \tag{5}$$

$$O_j = \sum_{i=1}^{m} |W_{ij}| \tag{6}$$

3.2 低ランク性とスパース性の併用

ここからは 3.1 節で導入した低ランク性とスパース性を併用して更新対象パラメータ数削減を行う手法について述べる．本稿は指定した安全性の差分プライバシを満たすことを目的としているため，プライベートな学習データを追加で参照すると追加コストとしてノイズを大きくするか安全性を下げる必要が出てしまう．そのため，追加コストが必要とならない範囲で更新対象パラメータを削減することで，安全性を下げることなくモデルの有用性を向上させる．3.1 節で述べた通り，低ランク性とスパース性のどちらかだけでは削減できる更新対象パラメータ数は限られているため，提案手法ではその両立を行う．提案

手法の概念図を図 3 に示す．

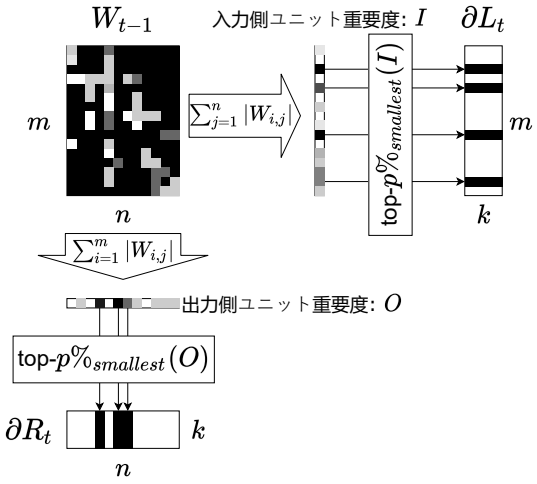


図 3　ユニットレベルの重要度を利用した
低ランク性とスパース性の併用

スパース性のみを利用した既存研究 [7] で用いられているシナプスの重要度 $\mathrm{abs}(W) \in \mathbb{R}^{m\times n}$ はランク分解された勾配行列 $\partial L \in \mathbb{R}^{m\times k}, \partial R \in \mathbb{R}^{k\times n}$ と対応が取れず，利用することが出来ない．そこで，∂L の行は入力側ユニットと，∂R の列は出力側ユニットに対応があることに注目する．式 (5)(6) で定義したユニットの重要度を利用し重要度の低い $p\%$ の重みに対応する勾配を 0 にすることでスパース化する．この処理は以下の式で表される．

$$\forall j \in \{1..k\}, \partial L_{ij} = \begin{cases} 0 & \text{if } I_i \text{ in top-}p\%_{smallest}(I) \\ \partial L_{ij} & \text{otherwise} \end{cases}$$

$$\forall i \in \{1..k\}, \partial R_{ij} = \begin{cases} 0 & \text{if } O_j \text{ in top-}p\%_{smallest}(O) \\ \partial R_{ij} & \text{otherwise} \end{cases}$$

これにより，更新対象パラメータ数は $k(m+n)\frac{100-p}{100} < k(m+n) \ll mn$ (where $k \ll \min(m,n) \wedge 0 < p < 100$) に削減できる．

なお，初期の重み W_0 をスクラッチから学習する場合は乱数による重みを，ファインチューニングを行う場合はパブリックデータで事前学習した重みを利用することで，ステップ 1 で追加のプライバシコストを必要としない．更に，ステップ t で参照する W_{t-1} は差分プライバシを満たす ∂W_{t-1} によって更新されているため，post-processing 定理 [2] により追加のプライバシコストを必要としない．まとめると，提案手法は，追加のプライバシコストを支払うことなくステップに応じた最新の情報に基づいたスパース性の利用となっている．これは，ステップ数に対して不変なパブリックデータで事前学習した重みに基づくスパース性を利用している既存研究 [7] と異なる点である．ファイチューニングが進むにつれて重要になる部分ネットワークは変化することが知られており [20]，提案手法の有用性向上に寄与していると考えられる．

最後に，詳細な手順を Algorithm 1 に示す．あるステップ t におけるあるレイヤ l のパラメータ更新をランク k，スパース性 p で選択的に行うことを考える．入力側出力側それぞれでユニットの重要度を算出し（2 - 7 行），重要でないユニットを選択する（8 - 9 行）．RGP や LoRA といった手法からランク分解された勾配行列を得た後（10 行），DP-SGD によるクリッピングとノイズ加算を行う（11 - 12 行）．このとき $'I^{(l)}$，$'O^{(l)}$ を勾配のノルム計算時にマスクとして利用する．その後スパース化を行い（13 - 22 行），最後に得られた勾配を利用して重みの更新を行う（23 - 24 行）．

Algorithm 1 低ランク性とスパース性を利用した
選択的パラメータ更新による差分プライベート学習
（ステップ t，レイヤ l）

Input: weights at previous step $W_{t-1}^{(l)} \in \mathbb{R}^{m\times n}$, layer $l \in H$, current step t, variance σ^2, clipping threshold C, rank k, sparsity p, external low-rank mechanism $LR(step, rank)$, DP mechanism $DP(gradients, clipping\_threshold, variance, mask)$, update mechanism $Update(previous\_weights, low\_rank\_gradients, low\_rank\_gradients)$

Output: weights at step t $W_t^{(l)}$

1: // $W_{t-1}^{(l)}$ is randomly initialized or pre-trained with public datasets or trained with private datasets by DP-SGD
2: **foreach** $i \in \{1..m\}$ **do**
3: $\quad I_i^{(l)} \leftarrow \sum_{j=1}^{n} |W_{t-1,(i,j)}^{(l)}|$
4: **end for**
5: **foreach** $j \in \{1..n\}$ **do**
6: $\quad O_j^{(l)} \leftarrow \sum_{i=1}^{m} |W_{t-1,(i,j)}^{(l)}|$
7: **end for**
8: Unimportant input units $\tilde{I}^{(l)} \leftarrow$ top-$p\%_{smallest}(I^{(l)})$
9: Unimportant output units $\tilde{O}^{(l)} \leftarrow$ top-$p\%_{smallest}(O^{(l)})$
10: $\partial L_t^{(l)}, \partial R_t^{(l)} \leftarrow LR(t,k)$ ▷ Use low-rankness
11: $\partial L_t^{(l)} \leftarrow DP(\partial L_t^{(l)}, C, \sigma^2, \tilde{I}^{(l)})$ ▷ Clip and Add noise
12: $\partial R_t^{(l)} \leftarrow DP(\partial R_t^{(l)}, C, \sigma^2, \tilde{O}^{(l)})$ ▷ Clip and Add noise
13: **foreach** $i \in \tilde{I}^{(l)}$ **do**
14: $\quad$ **foreach** $j \in \{1..k\}$ **do**
15: $\quad\quad \partial L_{t,(i,j)}^{(l)} \leftarrow 0$ ▷ Use sparsity
16: $\quad$ **end for**
17: **end for**
18: **foreach** $j \in \tilde{O}^{(l)}$ **do**
19: $\quad$ **foreach** $i \in \{1..k\}$ **do**
20: $\quad\quad \partial R_{t,(i,j)}^{(l)} \leftarrow 0$ ▷ Use sparsity
21: $\quad$ **end for**
22: **end for**
23: $W_t^{(l)} \leftarrow Update(W_{t-1}^{(l)}, \partial L_t^{(l)}, \partial R_t^{(l)})$
24: **return** $W_t^{(l)}$

4　評価実験

まず 4.1 節で総合的な評価を報告した上で，4.2 節でマイクロベンチマークとして低ランク性とスパース性の兼ね合いについて評価を報告し議論を行う．

タスク：ベンチマークには自然言語理解ベンチマークである

General Language Understanding Evaluation (GLUE) [9] を用いる．GLUE に複数設定されているタスクのうち，感情判定タスクである SST-2 と質問文対判定タスクである QNLI を対象とする．いずれも 2 値判定のタスクであるため，その正答率をモデルの有用性として評価する．なお，全ての実験で乱数シードのみを変更した 5 回分の結果の平均値を報告する．

モデル：自然言語処理タスクを扱うため，学習済み言語モデル RoBERTa [8] を用いる．事前学習データをパブリックなもの，追加学習データをプライベートなものとして扱った上でファインチューニングして評価を行う．RoBERTa として提供されているいくつかの学習済みモデルのうち，今回は Attention 層と全結合層を中心に構成され約 125 万パラメータから成る RoBERTa-base を利用する．

手法：提案手法として，(1) 低ランク化手法に RGP を利用したもの (Ours (RGP)) と (2) LoRA を利用したもの (Ours (LoRA)) の 2 つの実装を利用する．差分プライベート学習のベースラインとしては，(3) DP-SGD [4]，(4) RGP [5]，(5) LoRA+DP-SGD [6]，(6) Sparse DP-SGD の 4 つを利用する．DP-SGD は差分プライバシを満たす最も単純な手法，RGP と LoRA+DP-SGD は低ランク性を利用して DP-SGD の有用性を改善した手法，Sparse DP-SGD はスパース性を利用して DP-SGD の有用性を改善した手法である．Sparse DP-SGD は SNF-DP-SGD [7] と同じスパース性を利用するが，SNF-DP-SGD とは異なり，畳み込み層以外もスパース化の対象としている．提案手法，DP-SGD，RGP，Sparse DP-SGD はスクラッチからの学習とファインチューニングの両方に利用可能である一方で，LoRA+DP-SGD はファインチューニングに特化した手法である．また，差分プライバシを考慮しないベースラインとして，(7) N.P. を参考のため利用する．N.P. は単純なファインチューニングであり，差分プライベート学習の有用性上限の目安という位置付けである．

ハイパパラメータ：モデルに関するハイパパラメータは基本的に [6] に従う．バッチサイズは 2000，エポック数は 20，学習率は $1e-3$ とした．プライバシパラメータは $\delta = 1e-5$，クリッピングサイズ $C = 10$ とした．低ランク性を用いる既存手法である RGP と LoRA+DP-SGD ではそれぞれランク 1 と 16 で高い性能となることが報告されているが，本実験ではそれぞれ 8 と 32 がより高い正答率を示したため，性能を優先し，これらのパラメータを利用した評価を報告する．また，提案手法ではタスクと低ランク化手法ごとに表 1 のランクとスパース性を用いる．比較手法と同じく，正答率が高くなるようにパラメータを選択している．

表 1　提案手法で利用するランクとスパース性

タスク	低ランク化手法	ランク k	スパース性 p
SST-2	RGP	8	0.5
SST-2	LoRA	32	0.1
QNLI	RGP	8	0.1
QNLI	LoRA	32	0.1

4.1　総合的な評価

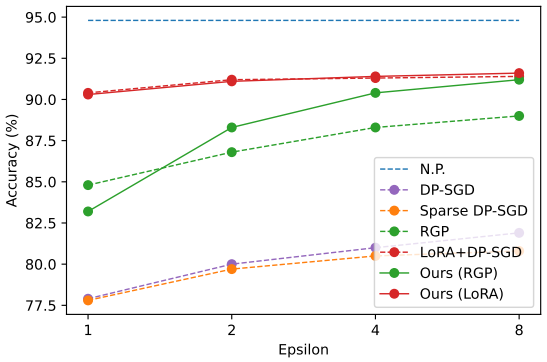
評価の指針とモチベーション：プライバシ強度パラメータ ϵ を $\{1, 2, 4, 8\}$ に変化させた際の各タスクの正答率から，手法ごとの総合的な性能を評価する．また，低ランク性またはスパース性のいずれかのみを利用した既存手法とそれらの併用を実現した提案手法を比較する．

結果：SST-2 タスクの実験結果を図 4(a) に，QNLI タスクの実験結果を図 4(b) に示す．低ランク性を導入することで最もナイーブな差分プライベート学習手法である DP-SGD から 5〜20%程度の性能向上が確認された．SNF-DP-SGD [7] に倣った Sparse DP-SGD は DP-SGD と比較すると，SST-2 タスクの場合は悪化，QNLI タスクの場合は改善と，タスクによって傾向が異なるものの，いずれも 1%未満の差であった．SNF-DP-SGD は畳み込み層のみを対象として性能改善をしていることから，畳み込み層と比較して RoBERTa の Attention 層や全結合層はあまりスパース性を仮定できない性質を持つ可能性が考えられる．SST-2 タスクの場合，提案手法のスパース性を取り入れた Ours (RGP) は RGP に対して正答率を 2%以上向上させることに成功したが，$\epsilon = 1$ の場合のみ逆に 1.5%程度低下した．これはスパース化による更新対象パラメータ削減が差分プライバシのためのノイズと比較して過剰になってしまったことによると考えられる．LoRA をベースとしている LoRA+DP-SGD と Ours (LoRA) はいずれも RGP をベースにした場合より全体的に高い性能を示した．特に ϵ が小さい場合でも性能低下が少ない傾向が特徴的である．LoRA+DP-SGD と Ours (LoRA) ではほぼ同等の性能を示した．これはスパース化によるクリッピングやノイズ影響の軽減と勾配情報の損失が釣り合う程度だったためだと考えられる．QNLI タスクの場合，低ランク性を利用した手法間の性能差は最大でも 1%程度と小さく，特にプライバシ強度が低いケースでは正答率がほぼ一致する形となった．これは，ノイズが少ない設定では，LoRA の学習済みの重みを固定して持ち続けることによるメリットが小さくなるためと考えられる．いずれの差分プライベートな手法もプライバシを考慮しない N.P. と比較すると 3%以上の性能低下が残る形となった．

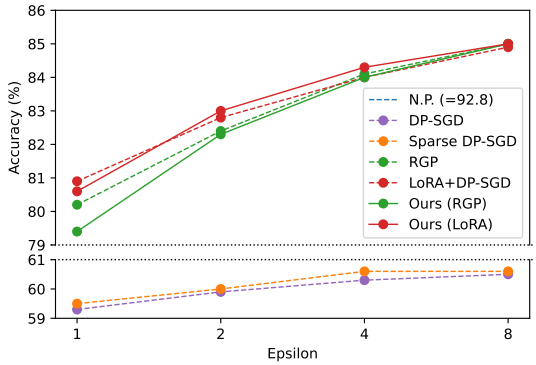
4.2　低ランク性とスパース性の併用に関する評価

評価の指針とモチベーション：ϵ を 2 に固定し，提案手法のパラメータであるランク k とスパース性 p を変化させた際の SST-2 タスクの正答率から，これらのパラメータの兼ね合いを確認する．また，ベースとなる低ランク性を取り入れる手法による傾向を確認する．

結果：低ランク性を利用する手法として RGP を利用した提案手法の実験結果を表 2 に，LoRA を利用した提案手法の実験結果を表 3 に示す．RGP を利用した場合，半数のパラメータに相当する $p = 0.5$ までスパース性を仮定してもランクに依らずほぼ単調に性能が向上した．$p = 0.7$ の場合，$k = 4$ でベスト性能になった一方，その他のランクでは性能が低下した．単純な削減パラメータ数だけでは性能の傾向が決まらないことがわかる．LoRA を利用した場合，一部のランクではスパース性



(a) SST-2 タスクの正答率



(b) QNLI タスクの正答率

図 4　GLUE ベンチマークによる RoBERTa の差分プライベートファインチューニングの評価．プライバシ強度パラメータ ϵ を変化させたときの正答率を扱う．実線は提案手法を，破線はベースライン手法を表す．また，同色の組み合わせは低ランク性のみを利用した既存手法とそれにスパース性を取り入れた提案手法の組み合わせを表す．なお，N.P. の結果は [8] に従う．

を仮定することで性能の向上が見られたが，ベスト性能を示したのはスパース性を仮定しないものだった．これはスパース化による勾配情報の損失がクリッピングやノイズ影響の軽減を上回ってしまったことによると考えられる．$p = 0.7$ のようにスパース性の仮定を強くおき，更にランクを小さくした場合，性能が大幅に低下することも確認された．これは過度な更新対象パラメータ数の削減により，ノイズ削減の効果以上に更新に必要な情報が失われてしまったことに起因すると考えられる．以上の結果から，ランクとスパース性の積で表現される削減パラメータ数だけでは単純な性能の傾向は決まらず，手法に合わせて適切なランクとスパース性を仮定する必要があることが示唆される．

3：スパース性の仮定がなく，RGP の結果に等しい．
4：スパース性の仮定がなく，LoRA+DP-SGD の結果に等しい．

表 2　$\epsilon = 2$ の SST-2 タスクでランクとスパース性を変化させたときの RGP を利用した提案手法の正答率 (%)

スパース性 p	ランク k = 2	4	8	16
0[3]	83.9	84.4	86.8	86.3
0.1	84.9	86.2	86.7	86.0
0.3	86.0	87.4	87.9	86.8
0.5	87.2	87.7	88.3	86.4
0.7	86.5	**88.7**	86.3	85.3

表 3　$\epsilon = 2$ の SST-2 タスクでランクとスパース性を変化させたときの LoRA を利用した提案手法の正答率 (%)

スパース性 p	ランク k = 8	16	32	64
0[4]	90.7	91.0	**91.2**	90.2
0.1	90.5	91.1	91.1	90.5
0.3	90.2	90.7	90.9	90.6
0.5	81.8	90.0	90.6	90.8
0.7	50.9	74.2	89.8	90.6

5　結　　論

本稿では，ニューラルネットワークパラメータの低ランク性とスパース性に基づき更新対象パラメータを適切に選択することで，差分プライバシを満たし有用性の高いモデルを獲得する学習手法を提案した．ユニットレベルの重要度を定義することで，これまで排他的であった低ランク性とスパース性の併用を実現した．実装を行い GLUE ベンチマークで評価したところ，低ランク性を取り入れる手法として RGP を利用した場合プライバシ強度を変えることなく 2%以上の有用性向上が確認された一方で，LoRA を利用した場合は性能の変化はごく僅かであった．今後，未検証である畳み込み層を持つモデルやスクラッチからの学習での評価を行う予定である．また，提案手法は具体的な低ランク化手法に依らないため，今後差分プライベート学習のための新たな低ランク化手法が提案された際には，更なる性能向上への貢献が期待できる．

文　　献

[1] 寺田雅之．差分プライバシーとは何か．システム／制御／情報，Vol. 63, No. 2, pp. 58–63, 2019.
[2] Cynthia Dwork. Differential Privacy. *International Colloquium on Automata, Languages, and Programming*, pp. 1–12, 2006.
[3] Apple. Differential Privacy Overview. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf, 2016.
[4] Edgar Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew Myers, Shai Halevi, Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.
[5] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-

Yan Liu. Large Scale Private Learning via Low-rank Reparametrization. *Proceedings of Machine Learning Research*, 2021.

[6] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially Private Fine-tuning of Language Models. *arXiv*, 2021.

[7] Zelun Luo, Daniel J. Wu, Ehsan Adeli, and Li Fei-Fei. Scalable Differential Privacy with Sparse Network Finetuning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 00, pp. 5057–5066, 2021.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 2019.

[9] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[10] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical Low-Rank Gradient Compression for Distributed Optimization. *Conference on Neural Information Processing Systems*, 2019.

[11] Jian Xue, Jinyu Li, and Yifan Gong. Restructuring of deep neural network acoustic models with singular value decomposition. *Interspeech*, pp. 2365–2369, 2013.

[12] Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On Compressing Deep Models by Low Rank and Sparse Decomposition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 67–76, 2017.

[13] Sridhar Swaminathan, Deepak Garg, Rajkumar Kannan, and Frederic Andres. Sparse low rank factorization for deep neural network compression. *Neurocomputing*, Vol. 398, pp. 185–196, 2020.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv*, 2021.

[15] Song Han, Jeff Pool, John Tran, and William J Dally. Learning both Weights and Connections for Efficient Neural Networks. *Conference on Neural Information Processing Systems*, 2015.

[16] Jose M Alvarez and Mathieu Salzmann. Learning the Number of Neurons in Deep Networks. *Conference on Neural Information Processing Systems*, 2016.

[17] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *The International Conference on Learning Representations Workshop*, 2017.

[18] Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *The International Conference on Learning Representations*, 2018.

[19] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-Specific Batch Normalization for Unsupervised Domain Adaptation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 00, pp. 7346–7354, 2019.

[20] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The Lottery Ticket Hypothesis for Pre-trained BERT Networks. *Conference on Neural Information Processing Systems*, 2020.

[21] Chen Liang, Simiao Zuo, Minshuo Chen, Haoming Jiang, Xiaodong Liu, Pengcheng He, Tuo Zhao, and Weizhu Chen. Super Tickets in Pre-Trained Language Models: From Model Compression to Improving Generalization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2021.